DOCUMENT RESUME

ED 463 324 TM 033 756

AUTHOR Sireci, Stephen G.; Rizavi, Saba
TITLE Comparing Computerized and Human Scoring of Students' Essays.
INSTITUTION Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
SPONS AGENCY College Board, New York, NY.
REPORT NO No-354
PUB DATE 2000-00-00
NOTE 29p.
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS College Students; Comparative Analysis; Computer Uses in Education; *Essay Tests; Evaluators; Higher Education; *Interrater Reliability; *Scoring; *Test Scoring Machines
IDENTIFIERS *Percent of Agreement

ABSTRACT

Although computer-based testing is becoming popular, many of these tests are limited to the use of selected-response item formats due to the difficulty in mechanically scoring constructed-response items. This limitation is unfortunate because many constructs, such as writing proficiency, can be measured more directly using items that require examinees to produce a response. Therefore, computerized scoring of essays and other constructed response items is an important area of research. This study compared computerized scoring of essays with the scores produced by two independent human graders. Data were essay scores for 931 students from 24 postsecondary institutions in Texas. Although high levels of computer-human congruence were observed, the human graders were more consistent with one another than the computer was with them. Statistical methods for evaluating computer-human congruence are presented. The case is made that the percentage agreement statistics that appear in the literature are insufficient for comparing the computerized and human scoring of constructed response items. In this study, scoring differences were most pronounced when researchers looked at the percentage of essays scored exactly the same, the percentage scored the same at specific score points, and the percentage of exact agreement corrected for chance. The implications for future research in this area are discussed. (Contains 11 tables, 2 figures, and 15 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 463 324

Comparing Computerized and Human Scoring of Students' Essays[1]

Stephen G. Sireci

University of Massachusetts Amherst

Saba Rizavi

Educational Testing Service

U.S. DEPARTMENT OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

S. Rizavi

TM033756

[1] Laboratory of Psychometric and Evaluative Research No. 354. School of Education, University of Massachusetts, Amherst, MA. This research was commissioned by the College Board. However, the views and conclusions expressed in this paper are those of the authors and should not be interpreted as official positions of the College Board.

BEST COPY AVAILABLE

Abstract

Although computer-based testing is becoming popular, many of these tests are limited to the use of selected-response item formats due to the difficulty in mechanically scoring constructed-response items. This limitation is unfortunate because many constructs, such as writing proficiency, can be measured more directly using items that require examinees to produce a response. Therefore, computerized scoring of essays and other constructed response items is an important area of research. In this study, we compared computerized scoring of essays with the scores based on two independent human graders. Although high levels of computer-human congruence were observed, the human graders were more consistent with one another than the computer was with them. Statistical methods for evaluating computer-human congruence are presented. We argue that the "percentage agreement" statistics that appear in the literature are insufficient for comparing computerized and human scoring of constructed response items. In this study, scoring differences were most pronounced when looking at the percentage of essays scored exactly the same, the percentage scored the same at specific score points, and the percentage of exact agreement corrected for chance. The implications for future research in this area are discussed.

Key words: computer-based testing, computerized scoring, constructed-response items, grader reliability, inter-rater reliability, writing assessment.

Comparing Computerized and Human Scoring of Students' Essays

The use of computers to score essays is receiving increasing attention by psychometricians and testing agencies (Burstein Kukich, Wolff, Lu, & Chodorow, 1998; Page & Peterson, 1995; Page, Poggo & Keith, 1997). The economical benefits of computerized essay scoring are obvious. Hiring writing experts to score essays is expensive and time consuming. If computers can be used to score essays, test-takers can receive their scores quickly, and testing agencies would save time and money in training human graders and using them to score essays. In addition, computerized scoring is efficient in that it allows for immediate score reporting for examinees who take a test on a computer. Although computerized scoring of essays and other constructed-response items is attractive, it is difficult for many of us to imagine how a computer could "read" something as subjective as an essay, follow a scoring rubric, and provide a reliable and valid score for a test taker.

Advances in computer technology, such as computational linguistics, have made computerized essay scoring possible. Computerized essay scoring algorithms analyze multiple essay characteristics to grade essays in a manner consistent with pre-specified scoring rubrics. For example, Burstein et al. (1998) describe an algorithm that comprises "more than 60 variables that might be viewed as evidence that an essay exhibits writing characteristics described in the … essay scoring guide" (p. 2).

A review of published literature in this area found three computerized essay scoring programs that claimed success in using the computer to score essays[1]. Page and his colleagues (Page & Peterson, 1995; Page, Poggo & Keith, 1997) concluded the "Project Essay Grade" (PEG) system provides scores comparable to that of human raters. Burstein et al. (1998) reported similar results for "E-rater," which is the computerized essay scoring program used by

Educational Testing Service to score writing samples associated with the Graduate Management Admissions Test (GMAT). Larkey (Studies on PEG and E-rater suggest these computer programs produce essay scores that are very similar to those provided by human graders. For example, Page and Petersen (1995) reported that the Pearson correlation between PEG-derived essay scores and scores based on a pair of human graders was higher than the correlation between two pairs of human graders (.82 versus .78). Burstein et al. (1998) found that scores provided by E-rater were congruent with scores provided by humans, but that this congruence was slightly below the level exhibited by two humans scoring the same essays. Looking across 13 essay prompts from the GMAT, Burstein et al. reported Pearson correlations ranging from .79 to .87 across human/computer comparisons, and from .82 to .89 across human/human comparisons. Thus, with respect to reliability, PEG and E-rater appear to provide scores that are consistent with scores provided by humans. The results of these studies suggest that, at a minimum, the computer could replace the necessity for a second human grader.

The College Board recently implemented the WritePlacer writing skills test. This test requires students to write an essay in response to a single prompt. Students' WritePlacer scores are used for placement purposes, such as deciding whether an incoming student needs remedial writing instruction. Computerized scoring of WritePlacer essays would be beneficial because students could be provided with scores much more quickly than if the essays were to be scored by humans, and because the scoring process would be much less expensive. In the present study, we investigate the congruence between human and computerized scoring of WritePlacer essays. Specifically, we evaluate a large sample of students' WritePlacer essays that were scored by a computerized essay scoring system called IntelliMetric (Vantage, 1998a) as well as by two human graders.

Description and Previous Research on Intellimetric

IntelliMetric is "an intellgent scoring system that emulates the process carried out by human scorers" (Vantage Technologies, 1998a, p 2.). The computer scoring algorithm is "trained" using previously (human) scored "marker" essays at each point along the score scale to "infer the rubric and the pooled judgments of human scorers" (Vantage Technologies, 1998a, p. 2). Although there are no published studies of IntelliMetric in the literature, four studies were conducted by its developer that bear on the reliability of IntelliMetric's essay scores (Vantage Technologies, 1998a; 1998b; 1999a; 1999b). One of these studies, (Vantage Technologies, 1999b), also addressed the validity of IntelliMetric scores. These four studies used two primary criteria for evaluating the congruence between essay scores assigned by IntelliMetric and essay scores assigned by human graders. One criterion was "percent agreement," which they (and Burstein et al., 1998) defined as scores that fall within one-point of one another. Because such scores do not reflect complete agreement, we use the term "percent adjacent" to describe this criterion. The second criterion for evaluating human/IntelliMetric congruence was the Pearson correlation between IntelliMetric and human scores. Due to the ordinal nature of the essay score scales, in this paper, we compute Spearman rank-order correlations to evaluate score congruence across human and IntelliMetric essay scores. We also compute Pearson correlations to compare the present results with those from other studies.

None of the Vantage Technologies Research Briefs reported the percentage of essays scored the same (percent exact) across human graders and IntelliMetric. However, the confusion matrix was provided for two of these studies (1998b; 1999a), which allowed for these statistics to be calculated. Across these two studies, the percentage exact statistics ranged from 62% to 64%. Unfortunately, data regarding the congruence across the human raters was not provided. The

percent adjacent statistics ranged from 95% to 100% across these four studies, which is higher than the percentages reported by Burstein et al. (1998). Across the four research briefs, the human/IntelliMetric Pearson correlations ranged from .78 to .90, which is consistent with previous research (Burstein et al., 1998; Ellis & Petersen, 1995; Ellis et al., 1997).

Data

The data for this study were essay scores for 931 students from 24 post-secondary institutions in Texas. Each student wrote one essay. The student essays were in response to one of two essay prompts: 464 students responded to essay prompt "1" and 467 students responded to essay prompt "2." The students did not have a choice regarding the prompt to which they responded. Each student's essay was scored by two human raters and the computer (IntelliMetric). A total of 36 human scorers were involved in grading the essays. Unfortunately, data regarding the specific scorers who graded each essay were not available. The two human scores provided for each essay were labeled "scorer 1" and "scorer 2," even though any of the 36 scorers could be scorer 1 or scorer 2 (but not both) for a given essay. Each essay was scored on a four-point scale, with a score of one signifying the lowest possible score. In the operational scoring of WritePlacer essays, when the two human scores differed by more than one point, the essay was sent to a third reader who resolved the discrepancy. There were eight such cases in the present data. In these cases, only the two original scores were considered in the analysis.

Results

Descriptive Statistics

The median, mean, and standard deviation of the IntelliMetric and human scores are presented in Table 1, for each essay prompt. The medians were the same across the three sets of scores for both prompts, which is not surprising given the short, four-point scale. The means

were similar for the human scorers. The mean differences across scorer 1 and scorer 2 were .01 for both prompts 1 and 2. Slightly larger mean differences were observed across the human/IntelliMetric comparisons. These differences ranged from .04 to .09; in all cases, the means associated with the computer scores were higher than the human scores. The essay scores provided by IntelliMetric also exhibited less variability, as can be seen from the relatively smaller standard deviations (.11 to .15 smaller than the human standard deviations). Bar charts summarizing the essay score distributions for each set of human graders and IntelliMetric are presented in Figures 1 (prompt 1) and 2 (prompt 2). The two sets of human graders appear to have similar score distributions for both prompts. IntelliMetric assigns more scores of "3" and fewer scores of "1," "2," and "4," relative to the human scorers, which explains the relatively larger mean and smaller standard deviation.

[Insert Table 1 Here]

[Insert Figures 1 and 2 Here]

Comparing Mean Differences Across Prompts and "Scorers"

The mean differences among the two sets of human graders and IntelliMetric, and the mean differences across essay prompts, were tested for statistical significance by conducting a two-way (grader-by-prompt) analysis of variance (ANOVA), with repeated measures on one factor (grader). The results of this analysis are summarized in Table 2. The only statistically significant effect was among graders ($F_{(2,1858)}$=12.01), but the associated effect size was small (.013). Follow-up post-hoc comparisons revealed that the statistical significance was due to the IntelliMetric scores being statistically significantly different from each of the two sets of human graders, but again the effect sizes were small (.02, for both scorers 1 and 2). The prompt effect and the interaction were not significant, which indicates that the essay prompt to which the

students responded did not have an effect on their scores, and that the differences noted between the human and IntelliMetric scores were consistent across prompts.

[Insert Table 2 Here]

Correlations among "graders"

In addition to the descriptive statistics reported in Table 1, correlations among the IntelliMetric and the two sets of human graders were also computed. These correlations provide a relative index of the similarities among IntelliMetric and the human graders. Because IntelliMetric is designed to provide scores similar to an "average" reader, the correlation between IntelliMetric and the average of the two graders' scores is also provided. Due to the ordinal nature of these data, Spearman correlations were computed. These correlations are reported separately for each essay topic in Table 3. Pearson correlations were also computed so that the results could be compared to those obtained from previous studies. The Spearman correlations across the two human graders were .75 and .79 for essay 1 and essay 2, respectively. The human/ IntelliMetric correlations were noticeably lower, ranging from .62 to .67. The average human/ IntelliMetric Spearman correlation was .63 for essay 2, which was .16 lower than the Spearman correlation between the two sets of human-graded scores for the same essay (r_S=.79). The Pearson correlations, also displayed in Table 3, are slightly lower for the Human/Human comparison and slightly higher for the Human/IntelliMetric comparisons than the Spearman correlations, but they show the same general pattern. The correlations among the human scores are larger than the correlations involving human and computer scores.

[Insert Table 3 Here]

Congruence Among IntelliMetric and Human Scorers

Cross-tabulations of the scores provided by IntelliMetric and the human scorers are provided in Tables 4 through 9. In addition to computing correlations among scores, three additional indices are reported to gauge the consistency of scores provided by different graders: percentage of essays given the exact same scores (% exact), the percentage of essays scored within one-point of another (% adjacent), and kappa coefficient[2] (% agreement corrected for chance, Cohen, 1960). It is important to note that the essays were graded on a four-point scale in the present study, whereas the Burstein et al. (1998) and Page et al. (1995, 1997) studies involved essays scored on a six-point scale. Use of a shorter scale makes it more likely to obtain relatively higher % exact and % adjacent indices. Thus, the kappa coefficient and intra-class correlation indices (described below) are especially important measures of congruence for the present data.

[Insert Tables 4 through 9 Here]

The cross-tabulations of the scores provided by the two human scorers exhibited levels of congruence found in previous research (e.g., Auchter, Sireci, & Skaggs, 1993; Page & Peterson, 1995). The percentage of essays scored the same were 82% and 84% for prompts 1 and 2, respectively. The percentage of essays scored by humans that were within one-point of each other was 98% for both prompts. Across the 931 essays, there were 11 instances (1.2%) where one of the human graders assigned a score of 1 to an essay and the other human assigned a score of 3, and 5 (0.5%) instances where one human gave a score of 2 to an essay and the other human gave a score of 4. In 9 of the 11 cases where there was a "1 versus 3" discrepancy between the humans, IntelliMetric assigned a grade of 2. In the other two cases, IntelliMetric assigned a

grade of 3. In all of the 5 cases where there was a "2 versus 4" discrepancy between the human graders, IntelliMetric assigned a grade of 3.

The percentages of essays scored the same for the human/IntelliMetric comparisons ranged from 70.4% (scorer 2, prompt 2) to 77.6 (scorer 1, prompt 1). The % adjacent indices ranged from 99.4% to 100%. The lack of perfect "% adjacent" indices for IntelliMetric were due to two cases where IntelliMetric assigned a score of "3" to an essay that the one of the two human raters assigned a "1." In both cases, the other human rater assigned a score of "2" to the essay.

The kappa coefficients, also reported in Tables 4 through 9, mirror the results from the correlation analyses. The scores provided by the two sets of human graders were more similar to one another than were the scores between IntelliMetric and either set of human scores. For the human scorers, the kappa coefficients were .66 and .71 for prompts 1 and 2, respectively. For the IntelliMetric /human comparisons, the kappa coefficients ranged from .51 to .52.

Intra-class correlations

Intra-class correlations (Ebel, 1951) were also computed among the human and IntelliMetric scores. These correlations are more informative regarding scorer congruence than the Spearman and Pearson correlations reported earlier because the intra-class correlation is sensitive to differences in the patterns of scores from two graders. For example, if the relative ranking of students' scores were the same across two graders, a perfect Pearson correlation of 1.0 could be obtained even if all scores across the graders differed by a constant (e.g., one grader assigns only scores one through three and the other grader assigns only scores two through four). On the other hand, the intra-class correlation reaches its maximum value of 1.0 only when 100% agreement occurs between two graders' scores. For this reason, the intra-class correlation is a

popular measure of inter-rater reliability. However, it should be noted that the intra-class correlation reflects the reliability of scores provided by a single grader. To estimate the inter-rater reliability of scores based on the average (or sum) of two graders, as is the case with WritePlacer essay scores, the intra-class correlations need to be adjusted using the Spearman-Brown formula. The Spearman-Brown adjusted intra-class correlation provides an estimate of the reliability of the essay scores based on an average, or sum, of scores provided by two or more independent readers.

The intra-class correlations among the sets of scores provided by humans and the computer were adjusted using the Spearman-Brown formula to provide an estimate of "grader reliability." These reliability estimates are presented in Table 10. Similar to the Pearson, kappa, and percent exact indices, greater congruence was observed among the human/human sets of scores (grader reliabilities ranged from .85 to .87) than among the IntelliMetric/human sets of scores (grader reliabilities ranged from .76 to .81).

[Insert Table 10 Here]

Inspection of Score Differences

Several results indicate that the IntelliMetric/human scoring discrepancies are due, in part, to the fact that IntelliMetric assigns more scores of 3 to the essays and fewer scores of 1 and 4, relative to the human graders (e.g. Figures 1 and 2, Tables 4 through 9). To explore the consistency of IntelliMetric and human scoring throughout the entire four-point score scale, conditional percentages were calculated. These conditional percentages indicate the percent of essays given a specific score by one grader (either IntelliMetric or a human) that were given the same score by each other grader. To calculate these percentages, the scores for one or more of the graders must be taken as the baseline for comparison. We do not know the "true" essay

scores, so we used each grader as the baseline and calculated separate conditional percentages for the other two graders. These conditional "percent exact" statistics are reported in Table 11. These percentages indicate that the only instance where IntelliMetric exhibited high agreement (i.e., >65%) with a human grader was when a human assigned a score of 3 to an essay. For example, of the 25 essays assigned a score of 1 by Scorer 1, only 8 (29%) were also scored 1 by IntelliMetric (see Table 1). Scorer 2 assigned a score of 1 to 19 (68%) of these same essays. The lowest conditional percentage agreements across human graders were for essays that were assigned a score of 4 by one of the two humans (range 36% to 59% exact). However, IntelliMetric was less likely to assign a grade of 4 to an essay that one or both of the humans scored as a four (range 13% to 36% exact).

[Insert Table 11 Here]

The data presented in Table 11 can be summarized by looking at the range of exact agreement indices at each essay score point. For a score of 1, the percent exact statistics range from 54% to 74% across human scorers and from 29% to 37% across human/IntelliMetric comparisons (where a human is used as the baseline). For a score of 2, the human exact agreement range is 74% to 85%. This range drops to 56% to 65% for human/IntelliMetric comparisons. For a score of 3, the percent exact statistics ranged from 86% to 92% across human graders and from 88% to 93% across the human/IntelliMetric comparisons. For a score of 4, the percent exact statistics range from 36% to 59% for humans, and from 13% to 36% for human/IntelliMetric comparisons. These results indicate that IntelliMetric is more likely to provide a score of 3 to an essay, and less likely to assign scores of 1, 2, or 4, relative to human graders.

Discussion

This study found that human graders who scored WritePlacer essays were more consistent with one another than IntelliMetric was with them. This finding is similar to the results reported by Burstein et al. (1998) for the E-rater system; however, there are several differences between the studies. Burstein et al. reported percent adjacent statistics ranging from .87 to .92 across human/E-rater comparisons, and from .87 to .93 across human/human comparisons. In this study, the percent adjacent statistics for the computer/human correlations were virtually 100%. A notable difference between the studies is that the present study involved essays that were scored on a four-point scale, whereas the essays in Burstein et al. were scored on a six-point scale. If all essays were assigned one of the two scale midpoints by a computer on a four-point scale (i.e., all 2 or all 3), the expected percentage adjacent due to chance would be 75%. If all of the essays were assigned to one of the two scale midpoints on a six-point scale (i.e., either all 3 or all 4), the expected percentage agreement due to chance would be 50%. Thus, the percentage adjacent statistics is not comparable across studies that involve different scoring scales.

The kappa coefficient is a more useful statistic for evaluating essay score congruence because it corrects for chance agreement. In this study, the kappa coefficients for the human/computer comparisons (.66 to .71) were noticeably larger than those observed for the human/computer comparisons (.44 to .52). Unfortunately, previous research in this area (i.e., Burstein et al., Page & Petersen, 1995; Page et al., 1997) did not report kappa coefficients; therefore, the present findings cannot be compared to these studies. However, the lower kappas observed for the human/IntelliMetric comparisons illustrate a meaningful discrepancy between the way humans and IntelliMetric scored these essays. The human/IntelliMetric percent exact

statistics (71%to 77%) were also lower than those found for the human/human comparisons (82% to 84%). However, the percentage of essays scored the same by IntelliMetric and humans in this study was higher than the percentages found in two of the previous studies reported by Vantage (1998b, 1999a; 62% and 64%, respectively).

Pearson correlations among human and computer essay scores were provided in previous research. The average correlations reported by Burstein et al. (1998) ranged from .79 to .87 for the human/E-rater comparisons, and from .82 to .89 for the human/human comparisons. In the present study, both sets of correlations were lower, and the difference between the human/human correlations and the human/computer correlations was larger. The human/human correlations ranged from .74 to .76, (similar to Page & Petersen, 1995) and the human/IntelliMetric correlations ranged from .64 to .69. The average human/PEG correlation reported by Page and Petersen (1995) was .82. The lower overall magnitude of the Pearson correlations found in this study could be due to the shorter, four-point essay score scale. However, the lower correlations observed for the human/computer comparisons, relative to the human/human correlations, cannot be explained by the shorter score scale.

Future research should investigate differences across the score scales and how they affect human/computer consistency. However, the ultimate criterion for determining the proper number of points along the essay score scale should be related to the construct validity of the essay scores (e.g., would a six-point scale facilitate more accurate placement decisions than a four-point scale?).

There are several limitations of the present study. The most glaring limitation is that when discrepancies are noted among computer and human graders, we do not know which scores are "correct." For example, it could be that the essays given a score of "3" by the computer

really are "3's" and the humans were wrong to give them other scores. Alternatively, it could be the other way around. Perhaps the computer is assigning scores of "3" to essays that are really better or worse than the quality reflected by a "3." The data gathered for this study cannot resolve this problem. This study sheds light on the relative consistency or reliability of human and computer WritePlacer scores, but does not shed any light regarding the relative validity of the human and computerized scoring processes. Future research should follow up the results of this study by employing an expert writing committee to review each essay and reach consensus regarding their "true" scores. These scores could then be used as a criterion for evaluating the scores provided by the human and computer graders (Williamson et al., 1999).

A second limitation of this study is that no data were available regarding which graders scored which essays. It could be that some of the human graders were inconsistent with their peers and the computer. Removing such graders may affect the results. Other limitations of this study are that only two essay prompts were evaluated (there are 16 total WritePlacer prompts), and students provided only one writing sample. To provide a proper estimate of the reliability of WritePlacer scores, an extended study should be conducted where students write essays in response to at least two prompts, and several prompts are administered. This type of study would provide an estimate of essay score reliability (in addition to the grader reliability reported here) and would provide an improved estimate of measurement error due to prompt variability.

Although the present study does have limitations, the results show that the findings of previous research cannot be generalized to IntelliMetric nor to the WritePlacer program. With respect to Pearson correlations, the levels of congruence observed among human graders and PEG (Page & Peterson, 1995), and among human graders and E-rater (Burstein, et al., 1998), were noticeably higher than those observed among human graders and IntelliMetric. However,

previous research did not report more informative comparative statistics such as kappa coefficients and grader reliabilities. A fair comparison of these different computerized essay scoring programs would require scoring the same set of essays with all three programs, and applying the same scoring and data analysis procedures to the results. If differences across the scores provided by these programs were observed in such a study, the differences among the computerized scoring algorithms underlying each of these programs should be compared and contrasted. Another important area of future research is the ability of computerized scoring algorithms to catch "fake" essays, or poorly-written essays that are written to receive a high score from a computer (e.g., an essay with a high word count that lacks content structure). Investigation of this issue would shed light on the "coachability" of writing samples that are known to be computer-scored.

References

Auchter, J. C., Sireci, S. G., & Skaggs, G. (1993). The tests of General Educational Development technical manual. Washington, D.C.: American Council on Education.

Burstein, J., Kukich, K., Wolff, S., Lu, C, & Chodorrow, M. (1998, April). Computer analysis of essays. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 10, 37-46.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychmetrika. 16(4), 407-424.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. Discourse Processes, 25 (2-3), 259-284.

Landauer, T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent semantic analysis and humans. In G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412-417). Mahwah, NJ: Lawrence Erlbaum Associates.

Larkey, L. S. (1998) Automatic Essay Grading Using Text Categorization Techniques. In Proceedings of the 21st Annual Internation Conference on Research and Development in Information Retrieval (SIGIR 98), Melbourne, Australia, pp. 90-95.

Page, B. P. & Peterson, N. S. (1995). The computer moves into essay grading: updating the ancient test. Phi Delta Kappan. 76(7), 561-565

Page, B. P., Poggo, J. P. & Keith, T. Z. (March, 1997). Computer analysis of student essays: finding trait differences in the student profile. Paper presented at the annual meeting of the Mid-South Educational Research Association. Chicago, IL (ERIC Document Reproduction Service No. Ed 411316).

SPSS Inc. (1998). SPSS Base 9.0: Applications Guide. Chicago: SPSS Marketing.

Vantage Technologies (1998a). How robust is IntelliMetric? A subsample cross validation study. Research Brief No. 301, Yardley, PA: Author.

Vantage Technologies (1998b). The efficacy of IntelliMetric for use in analyzing UK Secondary-level student narrative essays. Research Brief No. 304, Yardley, PA: Author.

Vantage Technologies (1999a). A study of IntelliMetric scoring of a seconday school admissions test requiring creative thinking. Research Brief No. 363, Yardley, PA: Author.

Vantage Technologies (1999b). Construct validity of IntelliMetric with international assessment. Research Brief No. 323, Yardley, PA: Author.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). 'Mental Model' comparison of automated and human scoring. Journal of Educational Measurement. 36 (2), 158-184.

Table 1: Medians, Means and Standard Deviations for Human and IntelliMetric Scores

Essay	Scorer-1			Scorer-2			IntelliMetric		
	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std
Prompt 1	3.0	2.65	0.64	3.0	2.64	0.68	3.0	2.73	0.53
Prompt 2	3.0	2.66	0.65	3.0	2.65	0.69	3.0	2.70	0.54

Table 2: Summary of ANOVA Results

Source	df	SS	MS	F	η^2
Essay Prompt	1	0.00	.00	0.01	.000
(b/w residual)	929	285.74	.31		
Graders	2	2.96	1.48	12.01*	.013
Graders X Prompt	2	0.31	.15	1.25	.001
(w/in residual)	1,858	228.07	.12		
Total	2,792	517.08			

*$p < .0001$

Table 3: Correlations Among Human and IntelliMetric Essay Scores

Spearman Correlations

Prompt	Scorer-1 & Scorer-2	Scorer-1 & IntelliMetric	Scorer-2 & IntelliMetric	Human Avg. & IntelliMetric
1	0.747	0.674	0.659	0.667
2	0.788	0.643	0.617	0.630

Pearson Correlations

Prompt	Scorer-1 & Scorer-2	Scorer-1 & IntelliMetric	Scorer-2 & IntelliMetric	Human Avg. & IntelliMetric
1	0.735	0.693	0.671	0.732
2	0.764	0.655	0.634	0.686

Table 4: Cross-tabulation for Scorer-1 versus Scorer 2 For Prompt 1

	Scorer 2				
Scorer 1	1	2	3	4	Total
1	**19**	6	3	0	28
2	10	**93**	15	2	120
3	3	27	**260**	12	302
4	0	0	6	**8**	14
Total	32	126	284	22	464

Percent exact: 82%
Percent adjacent: 98.3%
Kappa: .66

Table 5: Cross-tabulation for Scorer-1 versus Scorer 2 For Prompt 2

	Scorer 2				
Scorer 1	1	2	3	4	Total
1	**14**	2	3	0	19
2	10	**121**	13	3	147
3	2	19	**241**	12	274
4	0	0	11	**16**	27
Total	26	142	268	31	467

Percent exact: 84%
Percent adjacent: 98.3%
Kappa: .71

Table 6: Cross-tabulation for Scorer-1 versus IntelliMetric For Prompt 1

	IntelliMetric				
Scorer 1	1	2	3	4	Total
1	**8**	20	0	0	28
2	3	**68**	49	0	120
3	0	21	**278**	3	302
4	0	0	9	**5**	14
Total	11	109	336	8	464

Percent exact: 77.4%
Percent adjacent: 100%
Kappa: .51

Table 7: Cross-tabulation for Scorer-1 versus IntelliMetric For Prompt 2

	IntelliMetric				
Scorer 1	1	2	3	4	Total
1	**7**	11	1	0	19
2	1	**95**	51	0	147
3	0	28	**242**	4	274
4	0	0	20	**7**	27
Total	8	134	314	11	467

Percent exact: 75.2%
Percent adjacent: 99.8%
Kappa: .52

Table 8: Cross-tabulation for Scorer-2 versus IntelliMetric For Prompt 1

	IntelliMetric				
Scorer 2	1	2	3	4	Total
1	**9**	20	3	0	32
2	2	**71**	53	0	126
3	0	18	**263**	3	284
4	0	0	17	**5**	22
Total	11	109	336	8	464

Percent exact: 75.0%
Percent adjacent: 99.4%
Kappa: .49

Table 9: Cross-tabulation for Scorer-2 versus IntelliMetric For Prompt 2

	IntelliMetric				
Scorer 2	1	2	3	4	Total
1	**8**	18	0	0	26
2	0	**86**	56	0	142
3	0	30	**231**	7	268
4	0	0	27	**4**	31
Total	8	134	314	11	467

Percent exact: 70.5%
Percent adjacent: 100%
Kappa: .44

Table 10: Grader Reliabilities

Essay	Scorer-1/Scorer-2	Scorer-1/IntelliMetric	Scorer-2/IntelliMetric
Prompt 1	0.846	0.810	0.787
Prompt 2	0.865	0.784	0.762

Table 11: Conditional Percent Agreement Statistics for Each Essay Score Category

Prompt 1

	Scorer 1 Baseline			Scorer 2 Baseline			IntelliMetric Baseline		
Essay Score	N	Scorer 2 % exact	IntelliMet. % exact	n	Scorer 1 % exact	IntelliMet. % exact	n	Scorer 1 % exact	Scorer 2 % exact
1	28	68	29	32	59	28	11	73	82
2	120	78	57	126	74	56	109	62	65
3	302	86	92	284	92	93	336	83	78
4	14	57	36	22	36	23	8	63	63

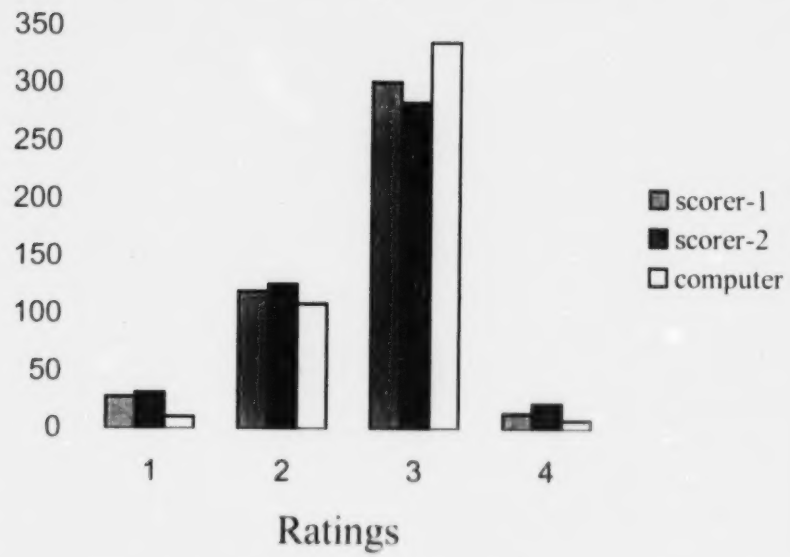
Prompt 2

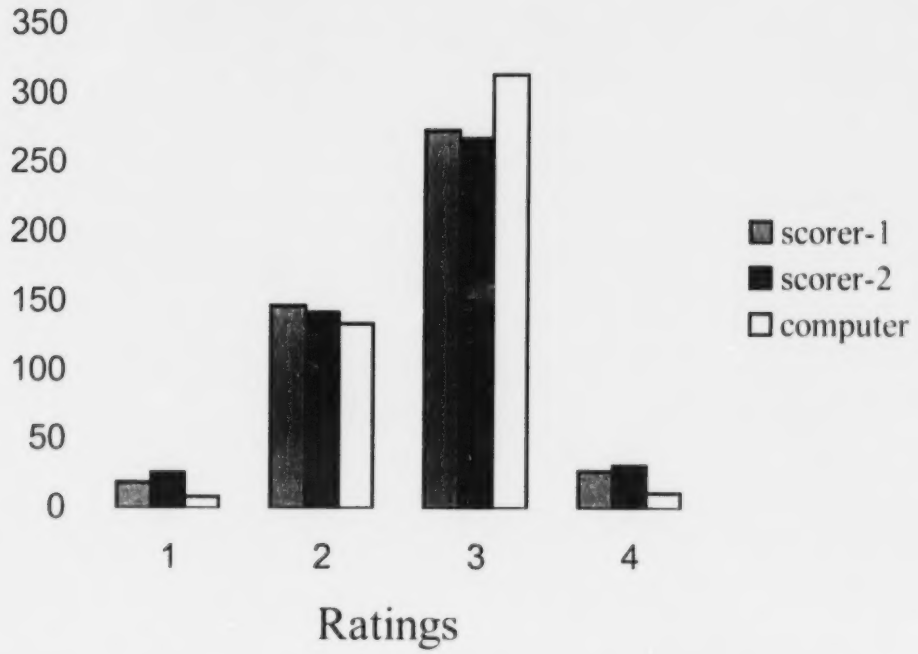
	Scorer 1 Baseline			Scorer 2 Baseline			IntelliMetric baseline		
Essay Score	N	Scorer 2 % exact	IntelliMet. % exact	n	Scorer 1 % exact	IntelliMet. % exact	n	Scorer 1 % exact	Scorer 2 % exact
1	19	74	37	26	54	31	8	88	100
2	147	82	65	142	85	61	134	71	64
3	274	88	88	268	90	88	314	77	74
4	27	59	26	31	52	13	11	64	36

Figure Captions

Figure 1. Comparison of Human and IntelliMetric Essay Score Distributions for Prompt 1

Figure 2. Comparison of Human and IntelliMetric Essay Score Distributions for Prompt 2


350
300
250
200
150
100
50
0
1
2
3
4
scorer-1
scorer-2
computer
Ratings


350
300
250
200
150
100
50
0
1
2
3
4
scorer-1
scorer-2
computer
Ratings

Footnotes

[1] The computer has also been shown to effectively score other performance tasks aside from essays. For example, Williamson, Bejar, & Hone (1999) describe computerized scoring of candidates' performance on the architecture licensing exam.

[2] The Kappa is a measure of inter-rater agreement that tests if the counts in the diagonal cells differ from those expected by chance alone. It is defined as, $K = (p_0 - p_e) / (1 - p_e)$, where p_0 = the sum of the observed proportions in the diagonal cells and p_e = the sum of the expected proportions in the same cells. The numerator is the excess beyond chance, and the denominator is the maximum that this value could be. When all off-diagonal cells are empty, Kappa obtains it's maximum value, 1.0. Values of Kappa greater than 0.75 indicate excellent agreement beyond chance, values between 0.40 to 0.75 indicate fair to good; and values below 0.40 indicate poor agreement (SPSS Inc., 1998).